REGRESSION PART 5: BIAS-VARIANCE TRAIDEOFF

Hsin-Min Lu
盧信銘
台大資管系

The Squared Loss Function (Section 1.5.5)

- Using RMSE or MSE as performance measure meaning that we are looking at squared loss function.
- Consider a scenario when we want to study the relationship between feature vector x and outcome t.
- We can think of x and t as random variables when we are collecting data points. They have a joint distribution $p(x, t)$.
- We want to construct our prediction function $y(x)$ by minimizing the expected squared loss:
- $E[L] = \int \int (y(x) - t)^2 p(x, t) dt dx$
- Meaning: minimize prediction error for all data.

The Squared Loss Function

- $E[L] = \iint (y(x) - t)^2 p(x,t)dtdx$
$$= \iint (y(x) - t)^2 p(t|x)p(x)dtdx$$
- Decompose $(y(x) - t)^2 = (y(x) - E[t|x] + E[t|x] - t)^2$
- $= (y(x) - E[t|x])^2 + 2(y(x) - E[t|x])(E[t|x] - t)$
$$+(E[t|x] - t)^2$$
- Taking expectation on the three terms separately
- $E[(y(x) - E[t|x])^2] = \int \int (y(x) - E[t|x])^2 p(t|x)p(x)dtdx$
$= \int (y(x) - E[t|x])^2 p(x)dx$ (why???)

The Squared Loss Function

- Recall $(y(x)-t)^2 = (y(x)-E[t|x])^2 + 2(y(x)-E[t|x])(E[t|x]-t) + (E[t|x]-t)^2$
- $E[2(y(x)-E[t|x])(E[t|x]-t)]$
  $$= \int\int 2(y(x)-E[t|x])(E[t|x]-t)p(t|x)p(x)dtdx$$
  $$= \int 2\,(y(x)-E[t|x])(E[t|x]-E[t|x])p(x)dx = 0$$
- $E[(E[t|x]-t)^2]$
  $$= \iint (E[t|x]-t)^2 p(t|x)p(x)dtdx = \int Var[t|x]p(x)dx$$

The Squared Loss Function

- Putting all three terms together
- $E[L] = \int (y(x) - E[t|x])^2 p(x)dx + 0 + \int Var[t|x]p(x)dx$
- The last term is the average of noise across all possible x.
  - We can do nothing about this term.
- We can minimize $E[L]$ by setting $y(x)$ to $E[t|x]$.
- That is, the best prediction that minimize the expected squared loss is $y(x) = E[t|x]$.
- If this is the case, then the expected square loss, or the expected MSE, is $\int Var[t|x]p(x)dx$

3.2 The Bias-Variance Decomposition (1)

- Recall the *expected squared loss*,
- $E[L] = \int (y(x) - E[t|x])^2 p(x)dx + \int Var[t|x]p(x)dx$
- The second term of $E[L]$ corresponds to the noise (variance) inherent in the random variable t. independent of the choice of y(x)
- Can we set $y(x) = E[t|x]$?
- In theory, yes, but…
- In reality, we do not know $E[t|x]$ for sure.

The Bias-Variance Decomposition (2)

- In reality, we are given limited dataset in order to learn $E[t|x]$.
- Following textbook notation, let $h(x) = E[t|x]$
- Suppose we were given multiple data sets, each of size N. Any particular data set, D, will give a particular function $y(x; D)$. We then have

$$\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

Cross product term

The Bias-Variance Decomposition (3)

- Taking the expectation over D, the cross product term vanished:
- $E_D[2\{y(x;D) - E_D[y(x;D)]\}\{E_D[y(x;D)] - h(x)\}] = 2\big[E_D[y(x;D)] - E_D[y(x;D)]\big]\{E_D[y(x;D)] - h(x)\} = 0$
- Thus we have:

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.$$

The Bias-Variance Decomposition (4)

- Putting everything together, we can write

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

- where

$$h(x) = E[t|x]$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\text{noise} = \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

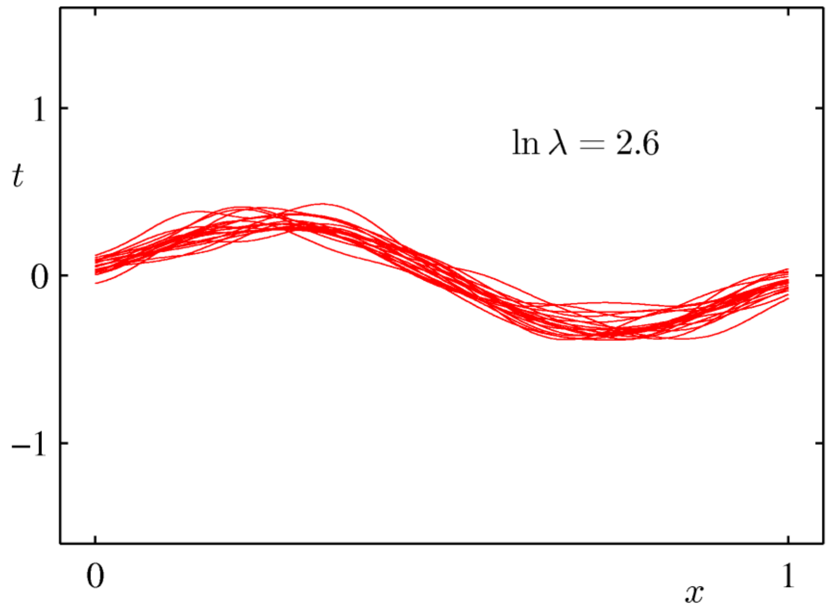
The Bias-Variance Decomposition (5a)

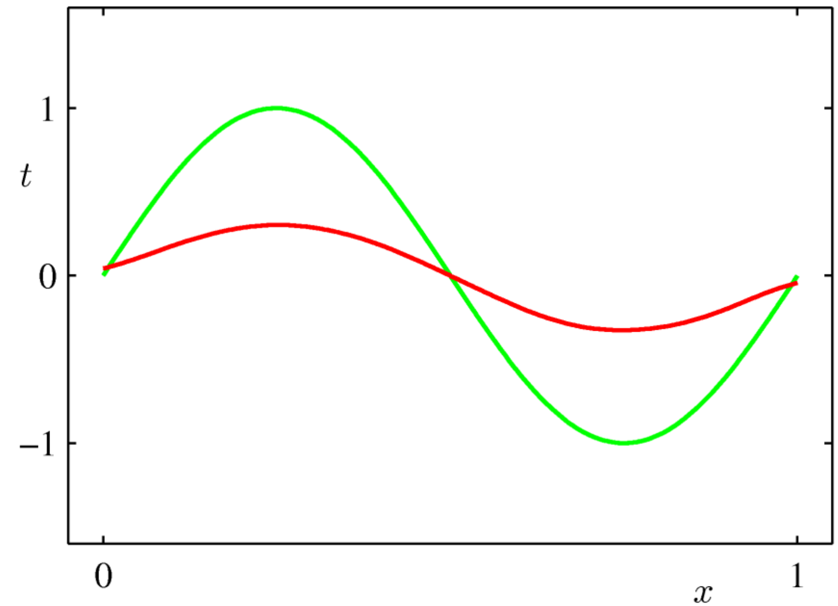
- Example: 100 data sets (each having N=25 data points) from the sinusoidal,
  varying the degree of regularization, λ.
  M=25, (24 Gaussian basis functions).

- Fitting data via minimizing $\frac{1}{2}\sum\left(t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n)\right)^2 + \frac{\lambda}{2}\sum w_j^2$
- Higher λ → More rigid model

The Bias-Variance Decomposition (5b)

- $\ln \lambda = 2.6$ (low variance, high bias)

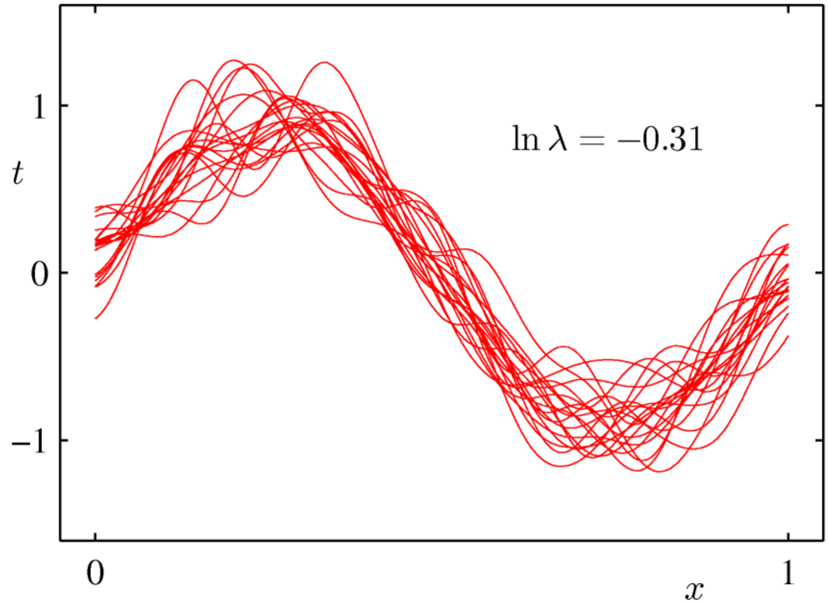


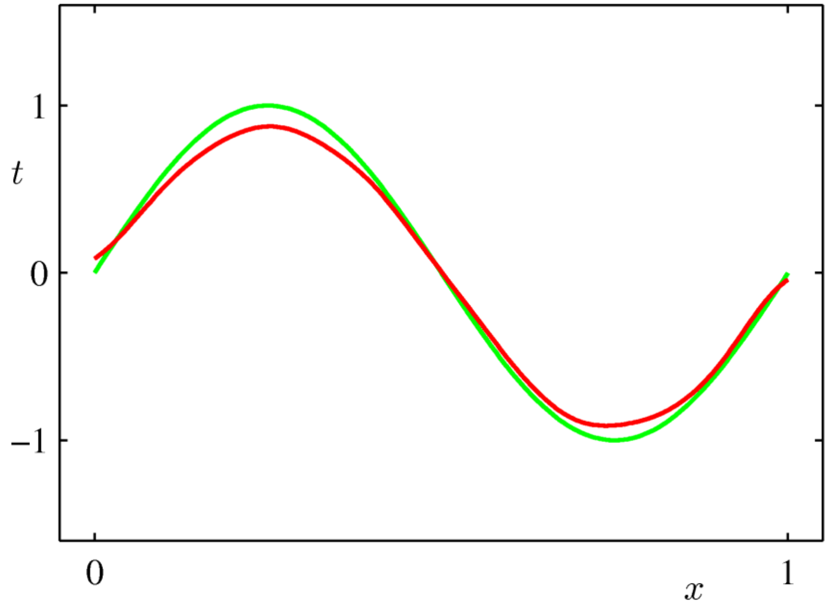


Green: true model
Red: average of 100 fits

The Bias-Variance Decomposition (6)

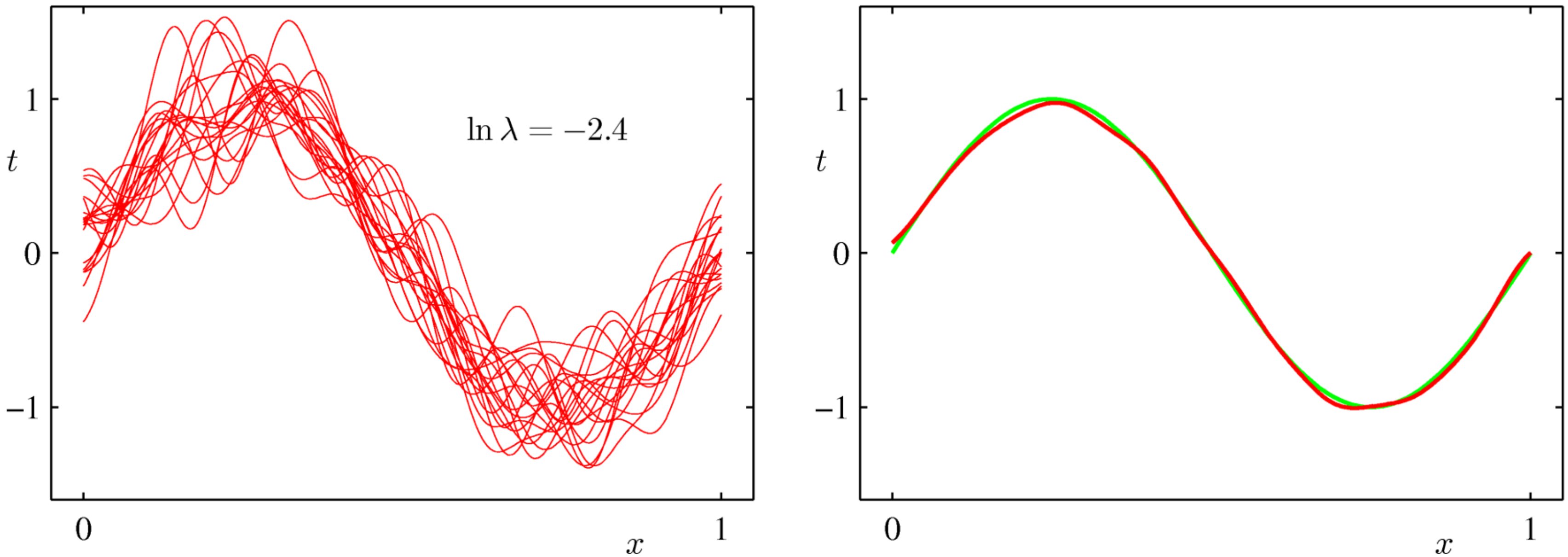
- $\ln \lambda = -0.31$ (smaller regularization)
- Higher variance, lower bias





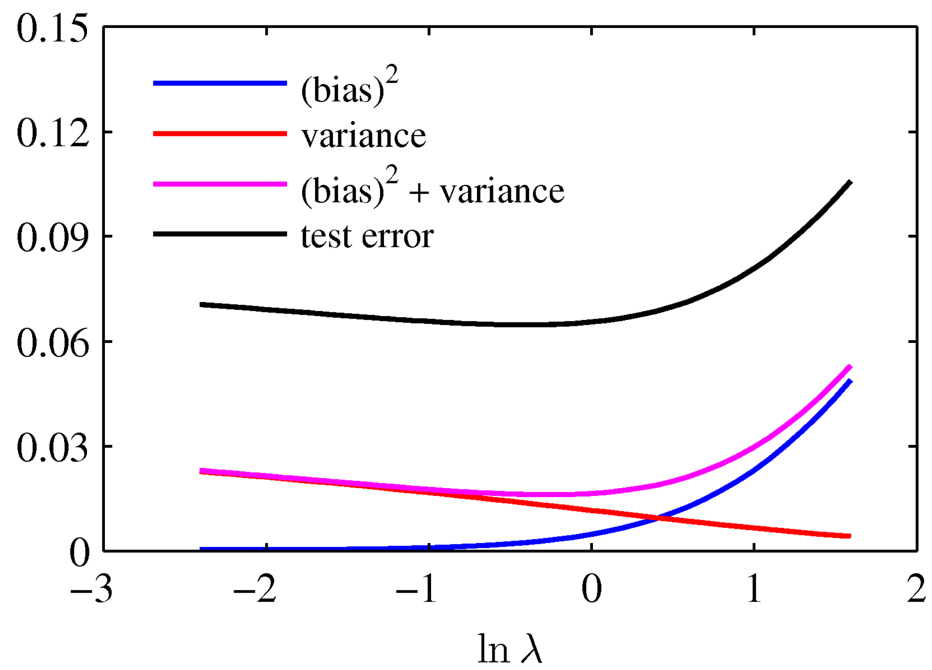
The Bias-Variance Decomposition (7)

- $\ln \lambda = -2.4$ (even smaller regularization)
- Even higher variance, even lower bias



The Bias-Variance Trade-off

- From these plots, we note that an over-regularized model (large λ) will have a high bias, while an under-regularized model (small λ) will have a high variance.
- We can usually gain some level of prediction accuracy by let-go unbiasedness and select a reasonable level of bias-variance trade-off.



Reading List

- PRML Ch 1.5.5, Ch 3.2